Nature Reviews Genetics | AOP, published online 10 November 2010; doi:10.1038/nrg2906

GENOMICS

A picture worth 1000 Genomes

Image created using Wordle (http://www.wordle.net) by C. Gunter, HudsonAlpha Institute for Biotechnology, Alabama, USA.

A cast of hundreds, if not quite thousands, of researchers worldwide have published their work on the pilot phase of the 1000 Genomes Project, building directly on the success of previous efforts of the Human Genome Project and the International HapMap Project. The 1000 Genome Project's stated goal is quite specific: "...to characterize over 95% of variants that are in genomic regions accessible to current high-throughput sequencing technologies and that have allele frequency of 1% or higher in each of five major population groups". In the pilot phase, the group tested three strategies for their relative yield towards this goal.

The group initially performed low-coverage (~2–6×) sequencing in 179 people from three populations. Then, to create a contrast with the data generated by such an approach, the group chose two extremes: deep-coverage (~42×) sequencing of one European-ancestry trio and one African trio, and exon sequencing of 8,140 exons in 697 individuals. As with most large-scale genomic projects, along the way the researchers came up with not just mountains of data (4.9 trillion bases in this case) but also innovations in storing and sharing the data, more accurate genotyping tools, and novel methods for alignment and analysis, among others.

> From the data, the authors were able to estimate that each person differs from the current reference genome by putative loss-of-function mutations in 250–300 genes.

As is clear from a Wordle picture generated from the paper's text, the main focus is on 'variants'. Using the above three approaches, the consortium was able to identify most SNPs that are already present in genomic databases. They also found some SNPs and many more structural variants that had not previously been described. From the data, the authors were able to estimate that each person differs from the current reference genome by putative loss-of-function mutations in 250–300 genes. Each of us is also heterozygous for 50–100 variants that are expected to cause inherited disorders, suggesting that genetic counselling based on whole-genome sequencing will continue to require much time and analysis (as well as the generation of a new definition of 'normal').

Another first from the project was a precise view of the patterns of selection acting on genic regions across multiple populations — a pattern that was visible even from the low-coverage sequencing data. Contrasting multiple human populations against the rhesus macaque reference genome as an outgroup demonstrates that, on average, diversity around a gene is reduced by 10% for distances up to 0.1 cM — typically 85 kb — from the gene. The authors also address a debate on recombination: based on the sequence variation around specific recombination hot spots, they infer no evidence for recombination itself creating a local mutagenic effect.

To use the data from the 1000 Genomes Project on your own burning questions, try using the Ensembl-based 1000 Genomes Browser.

Chris Gunter,
HudsonAlpha Institute for Biotechnology

The author declares no competing financial interests.

ORIGINAL RESEARCH PAPER The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010)

WEBSITES

1000 Genomes Project: http://www.1000genomes.org
The 1000 Genomes Browser: http://browser.1000genomes.org
Human Genome Project: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
International HapMap Project: http://hapmap.ncbi.nlm.nih.gov
***Nature* web focus on the 1000 Genomes Project:** http://www.nature.com/nature/focus/1000genomes/index.html